Intelligent data Analysis

Homework #2

Due Date: Tuesday, Feb 13th, 9:00 PM

Please submit all the answers in the form of a single pdf file.

Consider the attached data file for the tasks described below. This dataset is taken from the UC-Irvine Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/banknote+authentication>). The dataset has data derived from real and fake banknotes. Four different features are derived by processing the image of each banknote. The fifth column of the dataset contains the class label of 0 or 1 indicating whether the note is fake or real. Our goal is to build a classifier that can learn the model of a real banknote and then predict the class for a new banknote. The dataset contains 1,372 records.

Perform the following tasks:

1. (10) Load the dataset in MATLAB as an array named d1 and compute the covariance matrix of the data by using the command: “cm = cov(d1);”. Write all the information that you can infer by interpreting the values in the covariance matrix cm.
2. (10) Do a scatter plot of attribute-1 and attribute-2. Use different colors to mark the points from the two different classes. Write your interpretation of the separability of the two classes using attribute1 and attribute2, and also any other insights that you can obtain from this scatter plot. The following Matlab commands may accomplish the goal of obtaining the scatter plot:
   1. D10=d1(d1(:,5)==0,:)
   2. D11=d1(d1(:,5)==1,:)
   3. scatter(D10(:,1),D10(:,2),’k’)
   4. hold
   5. scatter(D11(:,1),D11(:,2),’r’)
3. (20) On the scatter plot obtained in Q#2 above, draw the axis-parallel boundaries, using your intuition, to suggest the boundaries that an efficient decision tree may draw. Show the resulting partitions in the form of a decision tree.
4. (10) From the dataset randomly select 800 records records for testing, 200 for validation, and the rest for testing. These selections MUST be random. Use a random number generator function to generate random numbers between 1 and 1372, and then choose records for training, validation, and testing. Show the code used for selecting the three different sets of records. Write the record numbers of rows that are chosen for your test set.
5. (20) Use the training set to learn decision trees (use fitctree command of Matlab) by varying the parameter that controls the minimum number of records in a leaf node. For this parameter use the values of 1, 2, 5, 10, 15, 20, 25, 30, . . , 50. For each generated tree find its performance (accuracy) on the test and validation sets. Plot this accuracy value against the parameter values for both the data sets. Choose that value of parameter which in your view does not overfit the data. Justify your choice of the parameter value and the corresponding decision tree. Show the selected tree graphically using Matlab’s view command. The role of the validation data set is to help you pick that decision tree which gives the best performance without overfitting the training data.
6. (10) Test the selected decision tree using the test portion of the dataset. Report accuracy, precision, and recall for the test dataset. Report the precision and recall values for both classes separately.
7. (20) Consider only the attribute-1 and attribute-2 of the dataset for predicting the class label. Train a perceptron using Matlab’s Train function that finds the best discriminating surface for this 2-D data space. Report the weight vector output by this function and also report the number of data points that are misclassified. Plot all the data points as done in the scatter plot in Q#2 above. Now plot the line representing the learned perceptron on this scatter plot. Mark the points that are misclassified.